7К

УДК 004.93+519.2

Капустий Б.Е.[1], Таянов В.А.[2]
[1] НУ «Львовская политехника», г. Львов, Украина
[2] ФМИ им. Г.В. Карпенко НАН Украины, г. Львов, Украина
vtayanov@ipm.lviv.ua

Влияние размера обучающей выборки на обобщающую способность метрических алгоритмов классификации

В работе предлагается поход, обеспечивающий оценку влияния уменьшения размера классов базы данных на уровень распознавания в случае использования *k*NN метрических классификаторов, а также дает возможность определения по данной выборке оптимального значения *k*. Проведено симулятивное моделирование результатов влияния уменьшения обучающей выборки на результаты распознавания. Полученные результаты могут быть использованы для дальнейшего формирования обучающей выборки и её коррекции.

Введение

На сегодняшний день метрические алгоритмы классификации являются одними из самых распространенных при проектировании практических целевых систем распознавания (СР). Среди таких систем часто встречаются биометрические СР. Для метрических алгоритмов классификации характерна простота настройки и достаточное быстродействие.

В целом разработка алгоритмов классификации является отдельной и сложной задачей [1-3]. Поскольку построение СР включает этапы генерации признаков, их селекции, построения классификаторов и их оптимизации в зависимости от выбранных признаков, формирования доверительного интервала принятия решения и определения его достаточного размера, формирования базы эталонных объектов и т.д. [1], [4-7], то в большинстве случаев используют метрические алгоритмы классификации [2], [3], [8]. Среди них чаще всего рассматривают правило ближайшего соседа (nearest neighbor, 1NN), реже – правило *k* ближайших соседей (*k* nearest neighbors, *k*NN) и совсем редко – взвешенный *k*NN классификатор [3]. Метод парзеновского окна и метод потенциальных функций в таких системах практически не используют. Одним из главных недостатков этих алгоритмов является то, что выборку необходимо сохранять полностью, а время распознавания прямо пропорционально длине обучающей выборки. Поэтому и возникает необходимость в разработке подходов, которые дали бы возможность эффективно уменьшить обучающую выборку так, чтобы при этом уровень распознавания был не менее заданного согласно техническому заданию на разработку СР. Для СР сокращение обучающей выборки означает уменьшение размерности классов базы данных. При этом необходимо обязательно провести стратификацию (stratification) классов. Итак, далее установим, как влияет уменьшение размерности классов на достоверность классификации при помощи *k*NN алгоритма.

Кроме того, что ожидаемые результаты моделирования пригодны для оценивания влияния размера обучающей выборки на результаты распознавания, они также могут использоваться для формирования такой обучающей выборки, которая бы по-

зволила минимизировать эффект переобучения. Данный поход позволяет провести предварительную оценку выборки, её потенциальных возможностей обучения и корректности. После его предварительной апробации необходимо осуществить исключение малоинформативных и искажающих объектов из обучающей выборки.

Постановка задачи. Пусть X – пространство объектов (object space); Y – множество имен классов (class name set); $y^* : X \to Y$ – целевая функция (target function), значения которой известны лишь на объектах конечной обучающей выборки длины l: $X^l = (x_i, y_i)_{i=1}^l \subset X \times Y$, $y_i = y^*(x_i)$ [3]. В базе данных существуют классы эталонов (class patterns) C_i, $i = \overline{1, n}$, причём $s_i = |C_i|$ – размеры классов. Предполагается, что размеры s_i всех классов одинаковые и равны s. Поскольку существует выборка контрольных образов U, подающихся на распознавание, то общее количество образов, принимающих участие в процессе распознавания, равно $n \times s + |U|$. Пусть оцененная частота ошибок (error frequency) алгоритма классификации $a = \mu(X^l)$ на обучающей выборке $X^l \subseteq X^L$: $\nu(a,U) = \frac{1}{|U|} \sum_{x \in U} [a(u) \neq y^*(u)]$, где запись $x \in U$ означает, что объект относится к контрольной последовательности, а запись $[a(u) \neq y^*(u)]$ должна пониматься как функция индикации несовпадения ответа, даваемого алгоритмом $a(u)$, и правильного ответа $y^*(u)$ для этого объекта. Задача состоит в оценивании величины $\tilde{\nu}(a,U) = \frac{1}{|U|} \sum_{x \in U} [\tilde{a}(u) \neq y^*(u)]$ при понижении информационного покрытия (information class coverage reduction) $|C_i|$ классов-эталонов, где $\tilde{a} = \mu(X^{\tilde{l}})$ – алгоритм, построенный на основании выборки размера $\tilde{l}$. В качестве алгоритма классификации используется алгоритм *k*NN. При такой общей постановке задачи наиболее пригодным подходом к её решению является комбинаторный подход. Очевидно, что в каждом конкретном случае понижение информационного покрытия классов (information class coverage reduction) может проводиться не обязательно оптимальным образом, однако общая статистика всех возможных понижений классов и результатов таких понижений должна дать ответ на вопрос об эффективности информационного покрытия классов-эталонов в целом.

Алгоритм ближайшего соседа. Представим данные, подающиеся на классификатор a, в виде двоичной последовательности $\{0,1\}$, посортированной по минимуму расстояний объектов базы данных от тестового объекта, где 1 ставятся в соответствие образам, поддерживающим правильное распознавание (образы своего класса), а 0 – образам, мешающим такому распознаванию (образы чужих классов). Пример такой последовательности подан на рис. 1.

$$\underbrace{1111111111}_{l}\ \underbrace{\underbrace{000}_{m_1}\underbrace{111}_{k_1}\underbrace{00}_{m_2}\underbrace{1111}_{k_2}\underbrace{000}_{m_3}\underbrace{11}_{k_3}\ldots\underbrace{000}_{m_n}\underbrace{1111}_{k_n}}_{\{k,m\}}\ldots$$

Рисунок 1 – Модель распознавания при задании начального размера класса в виде двоичной последовательности

Из приведенного рисунка видно, что последовательность образов, поддерживающих распознавание, имеет размерность $l + k = s$. Однако различные образы существенно отличаются друг от друга по возможностям этой поддержки. Действи-

тельно, при использовании 1NN классификатора удаление $l-1$ образов из класса-эталона не изменит результатов распознавания. С другой стороны, какой бы длинной ни была последовательность из k образов, она не сможет поддержать распознавание при отсутствии стратегической последовательности размером l и присутствии последовательности размером m.

При понижении размера обучающей выборки необходимо учитывать тот факт, что если последовательность размером l присутствует в начальном классе, то в классе с меньшим информационным покрытием s^* она может исчезнуть, и наоборот, если её не было, то может появиться, однако с меньшим размером l^*.

Рассмотрим возможности 1NN классификатора. Определяющим преимуществом этого классификатора является простота реализации, а к недостаткам можно отнести следующие [3]:

– неустойчивость к погрешностям, созданным выбросами в обучающей выборке (выбросом называют объект определенного класса, находящийся в окружении объектов чужих классов);
– полную зависимость алгоритма от метрики между объектами и отсутствие параметров для настройки по обучающей выборке методами скользящего контроля или иными.
– низкое качество классификации.

Несмотря на указанные недостатки, 1NN классификатор может иметь существенно лучшую устойчивость к эффекту понижения размера обучающей выборки. Это связано с тем, что данный классификатор менее чувствительный к размеру классов, чем *k*NN.

Итак, возможны два случая: начальное распознавание правильное либо неправильное, и необходимо определить вероятность его успешности после понижения размера обучающей выборки. То есть для первого случая необходимо определить вероятность того, что распознавание останется правильным, а для второго – вероятность перехода распознавания из категории неправильного в категорию правильного. Представим вероятность правильного распознавания при применении 1NN классификатора как отношение событий, поддерживающих успешное распознавание, к общему количеству событий:

$$P(k,l,s)=\begin{cases}\dfrac{C_s^{s^*}-C_k^{s^*}}{C_s^{s^*}}=1-\dfrac{k!(s-s^*)!}{s!(k-s^*)!}, k\ge s^*;\\ 1, \textit{в обратном случае.}\end{cases} \qquad (1)$$

При вычислении вероятностей (1) учтено, что если $k<s^*$ и начальное распознавание было правильным, то понижение размера обучающей выборки не приведет к ухудшению результатов распознавания, то есть $P(k<s^* \mid P(s)=1)=1$. Выражение (1) определяет вероятность того, что распознавание будет успешным независимо от того, каким образом был уменьшен размер обучающей выборки для своего и чужих классов. Таким образом, эта вероятность является оценкой сверху по отношению к точному (в смысле комбинаторики) значению вероятности правильного распознавания. Сам принцип оценок сверху вероятности успешного распознавания состоит в том, что вычисление точного значения соответствующей вероятности требует применения многошагового итерационного процесса.

Уточнить значение вероятности (1) можно путем введения еще одной оценки сверху вероятности того, что перед последовательностью $\{k_i\}$ после понижения размера обучающей выборки базы данных не будет находиться последовательность $\{\bigcup_j m_j\}, j<i$.

После исключения из модели (рис. 1) стратегической последовательности она трансформируется к такому виду:

$$\underbrace{\underbrace{000}_{m_1}\underbrace{111}_{k_1}\underbrace{00}_{m_2}\underbrace{111}_{k_2}\underbrace{1000}_{m_3}\underbrace{11}_{k_3}\ldots\underbrace{000}_{m_n}\underbrace{1111}_{k_n}}_{\{k,m\}}\ldots$$

Рисунок 2 – Модель распознавания в виде двоичной последовательности при $\{l\} \neq \varnothing$

Таким образом, задача сводится к определению вероятности успешного распознавания после понижения размера обучающей выборки для случаев, когда начальное распознавание было неправильным. Эти вероятности вычисляются n раз для пар последовательностей $\{m_i, k_i\}, i = \overline{1,n}$. Итак, на данном этапе исходной последовательностью из всех единиц будет последовательность размера k.

Определение. Показателем выживания подпоследовательности $\{m_i, k_i\}$ является вероятность того, что в результате всех возможных комбинаций вхождений объектов из этой подпоследовательности в другие в ней останется хотя бы один объект из исходной подпоследовательности. Указанную вероятность можно записать в виде:

$$\begin{cases} P(m_i, k_i, \{l\} = \varnothing) = \dfrac{C_{s-m_i}^{s^*}}{C_s^{s^*}}\left(1 - \dfrac{C_{k-k_i}^{s^*}}{C_s^{s^*}}\right), k - k_i \geq s^*, m - m_i \geq s^*; \\ 1, \textit{в обратном случае.} \end{cases} \quad (2)$$

Если все образы из своего класса в результате их сортировки по величине расстояний от тестового образа попали в пределы списка $\{m, k\}$, то выражение (2) определяет вероятность того, что в этом списке будут находиться такие образы из своего и чужих классов, при которых распознавание пройдёт успешно. Эта вероятность вычисляется рекурсивно-итерационным способом на основании подпоследовательностей $\{k_i, m_i\}$:

$$P\left(\{l\} \neq \varnothing, \{k_1\} \neq \varnothing, \{m_1\} = \varnothing\right) = P\left(\{l\} \neq \varnothing, \{k_1\} \neq \varnothing\right) P\left(\{m_1\} = \varnothing\right) =$$
$$= \frac{C_{s-m_1}^{s^*}}{C_{l+k}^{s^*}}\left(1 - \frac{C_{k-k_1}^{s^*}}{C_s^{s^*}}\right), k - k_1 \geq s^*, m - m_1 \geq s^*;$$
$$P\left(\{l\} \neq \varnothing, \{k_1\} \neq \varnothing, \{k_2\} \neq \varnothing, \{m_1\} = \varnothing, \{m_2\} = \varnothing\right) =$$
$$= P\left(\{l\} \neq \varnothing, \{k_1\} \neq \varnothing, \{k_2\} \neq \varnothing\right) P\left(\{m_1\} = \varnothing, \{m_2\} = \varnothing\right) =$$
$$= \frac{C_{s-m_1-m_2}^{s^*}}{C_s^{s^*}}\left(1 - \frac{C_{k-k_1-k_2}^{s^*}}{C_s^{s^*}}\right), k - (k_1 + k_2) \geq s^*, m - (m_1 + m_2) \geq s^*; \quad (3)$$

....

$$P\left(\{l\} \neq \varnothing, \{k_i\}_{i=1}^n \neq \varnothing, \{m_i\}_{i=1}^n = \varnothing\right) = P\left(\{l\} \neq \varnothing, \{k_i\}_{i=1}^n \neq \varnothing\right) P\left(\{m_i\}_{i=1}^n = \varnothing\right) =$$
$$= \frac{C_{s-\sum_i m_i\ 2}^{s^*}}{C_s^{s^*}}\left(1 - \frac{C_{k-\sum_i k_i}^{s^*}}{C_s^{s^*}}\right), k - \sum_i k_i \geq s^*, m - \sum_i m_i \geq s^*.$$

В формулах (3) значения n определяются условиями $s-l-\sum_i k_i \geq s^*$ та $s-\sum_i m_i \geq s^*$, поскольку все дальнейшие вероятности $P(\cdot)$ равны 1. Произведение всех вероятностей (3) является глобальной вероятностью правильного распознавания.

Алгоритм *k* ближайших соседей. Представим результаты распознавания подобно тому, как они были представлены для 1NN случая, то есть в виде двоичной последовательности. Пример такой последовательности показан на рис. 3.

$$\underbrace{1111111111}_{l_1}\underbrace{000}_{m_1}\underbrace{111111111}_{l_2}\underbrace{00000}_{m_2}\underbrace{111}_{l_3}\underbrace{00}_{m_3}\ldots$$

Рисунок 3 – Результаты распознавания в виде двоичной последовательности

При использовании *k*NN классификатора важно, чтобы среди k ближайших соседей было относительное либо абсолютное большинство образов своего класса среди других образов. Рассмотрим более простой случай, предусматривающий относительное большинство. Успешная работа *k*NN классификатора состоит в том, что для k ближайших соседей выполняется условие

$$\left|\bigcup_i \tilde{l}_i\right| > \left|\bigcup_i \tilde{m}_i\right|, i = 1,2,3\ldots, \tag{4}$$

где $\tilde{l}_i, \tilde{m}_i$ – группы, образующиеся после понижения информационного покрытия классов. Под группой понимается однородная последовательность элементов. В последовательность (рис. 3) входят образы всех классов, хотя в общем случае однозначного соответствия между количеством групп и количеством классов не существует. Если рассматривать лишь случай нечётных значений k в *k*NN классификаторе, то исключается неоднозначность классификации, наблюдаемая при чётных значениях k и равенстве голосов за различные классы.

Оценим эффект от понижения информационного покрытия классов при использовании *k*NN классификатора. Примем, что размеры всех суженных классов одинаковые и равны s^*. Для *k*NN классификатора, в отличие от 1NN, не имеет такого принципиального значения последовательность первых образов своего класса. Поэтому произвольную последовательность образов своего класса можно обозначить как l_i.

Рассмотрим сначала случай $s^* = ENT\left(\frac{k}{2}\right)+1$. Определим вероятности того, что среди последовательности образов своего класса заданной длины будут выбраны комбинации из s^* образов. Такие вероятности носят доверительный характер и характеризуют степень покрытия несжатого класса последовательностью из $\left|\bigcup_i l_i\right|$ образов, среди которых выбирается s^*. Кроме них, найдём также вероятности того, что не будут выбраны соответствующим способом определённые образы из чужих классов. Вероятность правильной работы *k*NN классификатора является произведением этих двух вероятностей.

Обозначим вероятность ошибочной классификации, обусловленной образами чужих классов, для соответствующих групп m_i как q_j:

$$q_1 = P\left(\inf\left(\left|\bigcup_i m_i\right|\right) \geq ENT\left(\frac{k}{2}\right)+1\right);$$

$$q_2 = P\left(\inf\left(\left|\bigcup_i m_i\right|\right) + |m_{i+1}| \geq ENT\left(\frac{k}{2}\right)+1\right);$$

$$q_3 = P\left(\inf\left(\left|\bigcup_i m_i\right|\right) + |m_{i+1}| + |m_{i+2}| \geq ENT\left(\frac{k}{2}\right)+1\right);...\qquad(5)$$

$$q_j = P\left(\inf\left(\left|\bigcup_i m_i\right|\right) + \left|\bigcup_j m_{i+j-1}\right| \geq ENT\left(\frac{k}{2}\right)+1\right);...$$

Вероятность q_j для каждого из значений доверительной вероятности равна

$$q_j = \frac{C^{s^*}_{\left|\bigcup_{i,j} m_{i+j-1}\right|}}{C^{s^*}_s}.\qquad(6)$$

Соответствующую доверительную вероятность можно представить в виде:

$$P_{q_j} = P(\bigcup_i l_i) = \frac{C^{s^*}_{\left|\bigcup_i l_i\right|}}{C^{s^*}_s}.\qquad(7)$$

Итак, вероятность успешной работы *k*NN классификатора равна

$$P_j\left(\bigcup_i l_i, \bigcup_{i,j} m_{i+j-1}\right) = P_{q_j}(1-q_j) = \frac{C^{s^*}_{\left|\bigcup_i l_i\right|}}{C^{s^*}_s}\left(1-\frac{C^{s^*}_{\left|\bigcup_{i,j} m_{i+j-1}\right|}}{C^{s^*}_s}\right) = C^{s^*}_{\left|\bigcup_i l_i\right|}\left(\frac{C^{s^*}_s - C^{s^*}_{\left|\bigcup_{i,j} m_{i+j-1}\right|}}{\left(C^{s^*}_s\right)^2}\right).\qquad(8)$$

При $q_0 = 0$ и $\left|\bigcup_i m_i\right| < ENT\left(\frac{k}{2}\right)+1$ эта вероятность составляет

$$P_0\left(\left|\bigcup_i m_i\right| < ENT\left(\frac{k}{2}\right)+1\right) = P_{q_0} = P(\bigcup_i l_i) = \frac{C^{s^*}_{\left|\bigcup_i l_i\right|}}{C^{s^*}_s}.\qquad(9)$$

Рассмотрим второй случай $ENT\left(\frac{k}{2}\right)+1 < s^*$ и определим для него вероятность ошибочной классификации, обусловленной образами чужих классов:

$$q_j = \frac{\sum\limits_{j=ENT\left(\frac{k}{2}\right)+1}^{s^*-1} C^j_{\left|\bigcup_{i,j} m_{i+j-1}\right|} C^{s^*-j}_{s-\left|\bigcup_{i,j} m_{i+j-1}\right|}}{C^{s^*}_s}, \left|\bigcup_{i,j} m_{i+j-1}\right| \geq ENT\left(\frac{k}{2}\right)+1.\qquad(10)$$

Вычислим доверительную вероятность для произвольной последовательности из образов своего класса:

$$P_{q_j} = \frac{\sum_{j=ENT\left(\frac{k}{2}\right)+1}^{s^*-1} C^j_{\left|\bigcup_i l_i\right|} C^{s^*-j}_{s-\left|\bigcup_i l_i\right|}}{C_s^{s^*}}. \qquad (11)$$

Вероятность успешного распознавания при применении *k*NN классификатора определяется произведением вероятности (11) и дополнения к вероятности (10).

$$P_j = P_{q_j}(1-q_j) = \frac{\sum_{j=ENT\left(\frac{k}{2}\right)+1}^{s^*-1} C^j_{\left|\bigcup_i l_i\right|} C^{s^*-j}_{s-\left|\bigcup_i l_i\right|}}{C_s^{s^*}} - \frac{\left(\sum_{j=ENT\left(\frac{k}{2}\right)+1}^{s^*-1} C^j_{\left|\bigcup_i l_i\right|} C^{s^*-j}_{s-\left|\bigcup_i l_i\right|}\right)\left(\sum_{j=ENT\left(\frac{k}{2}\right)+1}^{s^*-1} C^j_{\left|\bigcup_{i,j} m_{i+j-1}\right|} C^{s^*-j}_{s-\left|\bigcup_{i,j} m_{i+j-1}\right|}\right)}{\left(C_s^{s^*}\right)^2}. \qquad (12)$$

Эта вероятность для ошибки $q_0 = 0$ составляет:

$$P_0\left(\left|\bigcup_i m_i\right| < ENT\left(\frac{k}{2}\right)+1\right) = \frac{\sum_{j=ENT\left(\frac{k}{2}\right)+1}^{s^*-1} C^j_{\left|\bigcup_i l_i\right|} C^{s^*-j}_{s-\left|\bigcup_i l_i\right|}}{C_s^{s^*}}. \qquad (13)$$

Итак, при $ENT(\frac{k}{2})+1 = s^*$ вероятность правильного распознавания для *k*NN классификатора вычисляется по формуле (8), а при $ENT\left(\frac{k}{2}\right)+1 < s^*$ – по формуле (12).

Результаты симулятивного моделирования. Было проведено моделирование процесса распознавания с разными последовательностями образов своего и чужих классов для 1NN и *k*NN классификаторов в случае относительного большинства. Моделирование использовано для оценивания результатов работы системы распознавания лиц людей [9], [10]. В связи с этим начальный размер классов был принят равным 18.

На рис. 4, 5 представлены результаты моделирования влияния уменьшения размера обучающей выборки на вероятность правильного распознавания для 1NN классификатора. На рис. 4 показана зависимость доверительной вероятности правильного распознавания от размера последовательности образов своего класса и размера классов базы эталонов. Как видно из рисунка, доверительная вероятность уменьшается при уменьшении размера эталонных классов и последовательности образов своего класса. На рис. 5 изображена зависимость вероятности правильного распознавания от размера последовательностей образов своего и чужих классов в случае их попарного разделения. Моделирование проводилось следующим образом. Формировалась последовательность переменного размера из образов своего класса, а к ней периодически прибавлялось по одному образу из чужого класса, что привело к формированию совокупной переменной последовательности из образов своего и чужого классов. Для каж-

дой такой последовательности и соответствующего размера класса вычислялась вероятность правильного распознавания. Из рисунка видно, что увеличение последовательности из образов чужого класса приводит к уменьшению вероятности правильного распознавания.

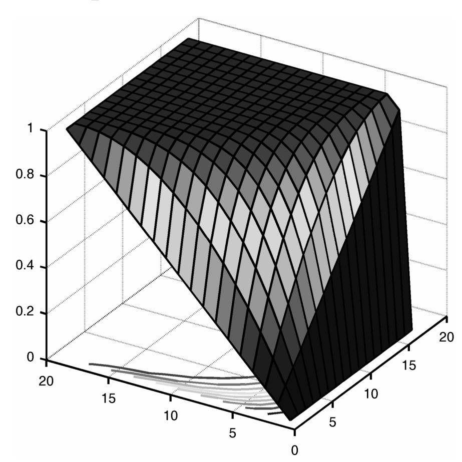


Рисунок 4 – Доверительная вероятность правильного распознавания как функция размера классов базы эталонов (ось x) и размера последовательности образов своего класса (ось y) для 1NN классификатора

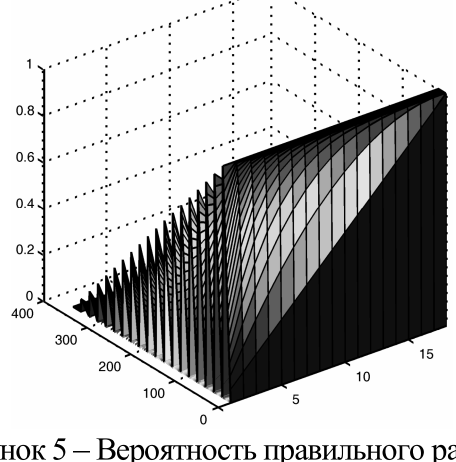


Рисунок 5 – Вероятность правильного распознавания как функция размера классов базы эталонов (ось x) и размера последовательности образов своего и чужих классов (ось y) для 1NN классификатора

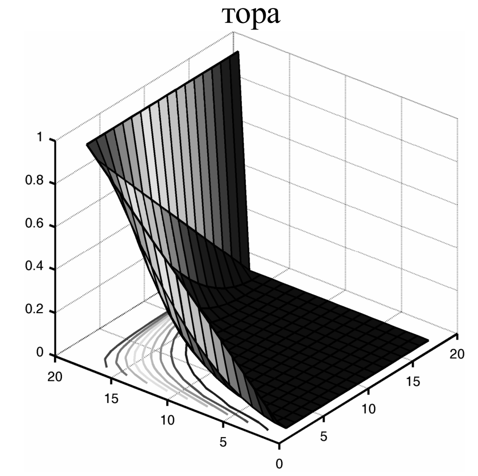


Рисунок 6 – Доверительная вероятность правильного распознавания как функция размера классов базы эталонов (ось x) и размера последовательности образов своего класса (ось y) для *k*NN классификатора при $s^* = ENT\left(\frac{k}{2}\right)+1$

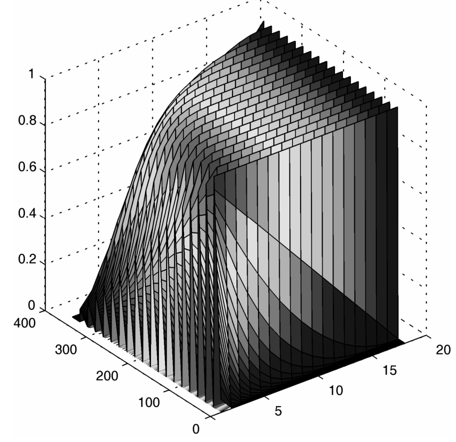


Рисунок 7 – Вероятность правильного распознавания как функция размера классов базы эталонов (ось x) и размера последовательности образов своего и чужих классов (ось y) для *k*NN классификатора при $s^* = ENT\left(\frac{k}{2}\right)+1$

На рис. 6, 7 представлены результаты моделирования влияния уменьшения размера обучающей выборки на вероятность правильного распознавания для kNN классификатора в случае, когда $ENT\left(\frac{k}{2}\right)+1 = s^*$.

На рис. 6 показана аналогичная к рис. 4 зависимость. Как видно из рисунка, доверительная вероятность уменьшается при увеличении числа ближайших соседей. Результаты, изображённые на рис. 7, указывают на то, что вероятность правильного распознавания уменьшается при увеличении размера класса и последовательности образов чужого класса.

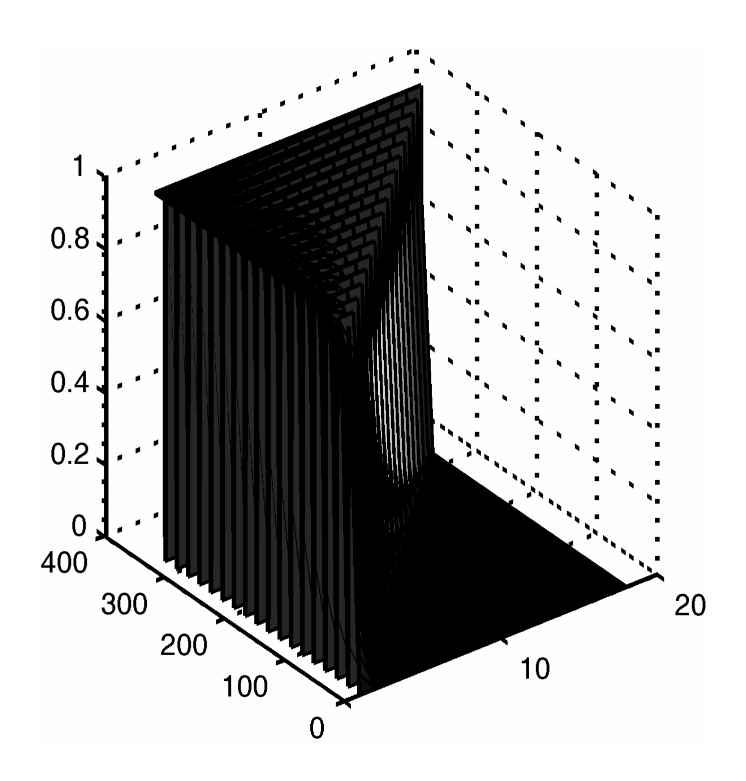


Рисунок 8 – Доверительная вероятность правильного распознавания как функция размера классов (ось x) и значения $ENT\left(\frac{k}{2}\right)+1$ (ось y)

На рис. 8, 9 приведены результаты моделирования для случая kNN классификатора и $ENT\left(\frac{k}{2}\right)+1<s^*$. На рис. 8 представлена зависимость доверительной вероятности от размера класса и значения $ENT\left(\frac{k}{2}\right)+1$. Размер класса периодически увеличивался, и также периодически формировалась переменная последовательность образов своего класса. Общая последовательность представлена одной из координат, а второй – значение $ENT\left(\frac{k}{2}\right)+1$. Зависимости на рис. 9 построены для таких случаев: $ENT\left(\frac{k}{2}\right)+1=\{1,3,5,7,9,11,13,15,17\}$. Вероятность правильного распознавания будет тем больше, чем меньше последовательность образов чужих классов и больше разница между $ENT\left(\frac{k}{2}\right)+1$ и s^*.

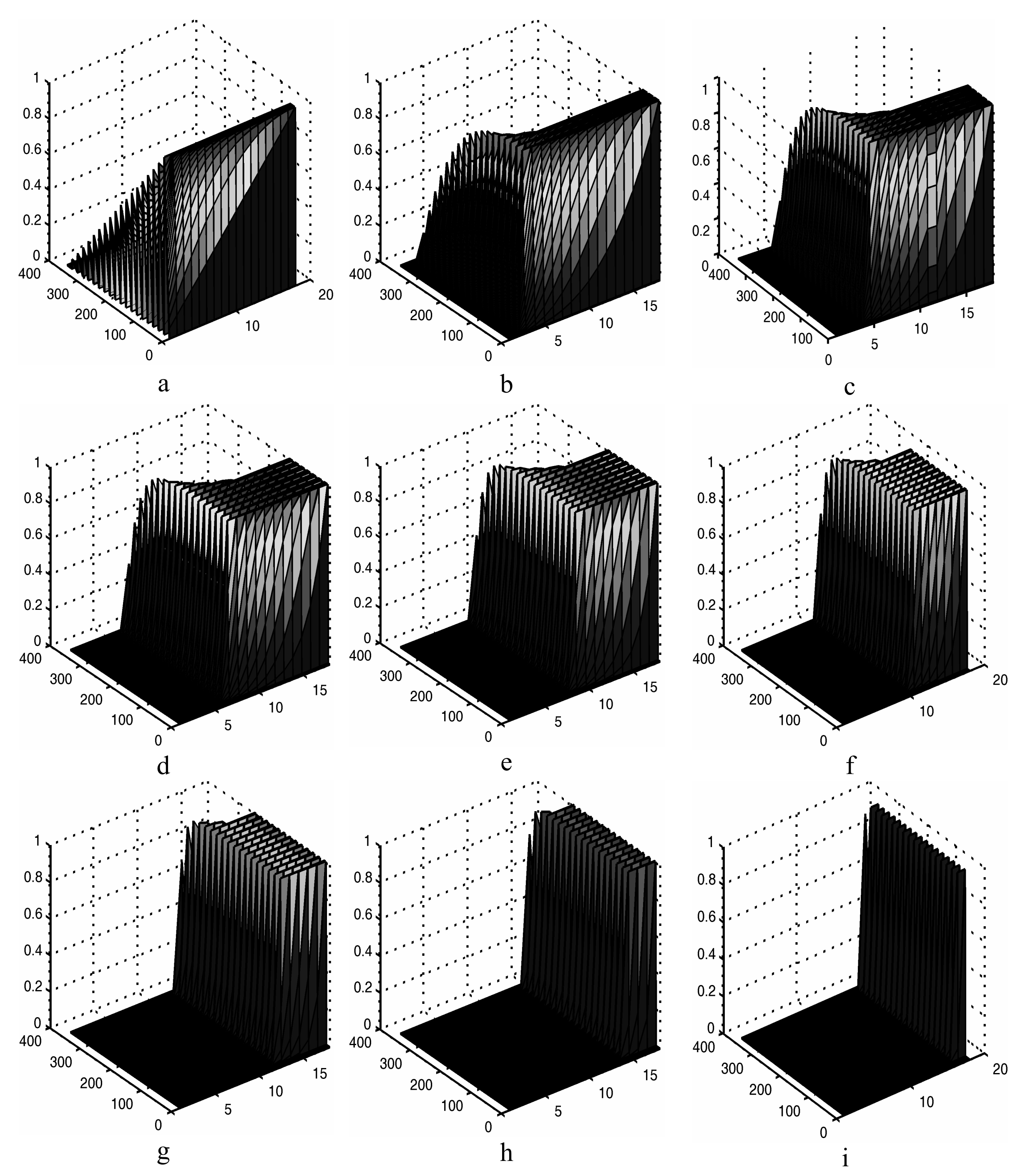


Рисунок 9 – Вероятность правильного распознавания как функция $ENT\left(\frac{k}{2}\right)+1$ (ось x) и общей последовательности образов своего и чужих и классов (ось y)

Выводы

На основании комбинаторного подхода можно анализировать и оптимизировать *k*NN классификатор. Подход даёт возможность определять соотношение между оптимальным значением k ближайших соседей и пониженным размером класса s^*. Это соотношение зависит от результатов распознавания на начальных (несжатых)

классах. Подход также даёт дополнительную информацию о составе обучающей выборки и степени её корректности для того, чтобы в дальнейшем использовать эту информацию при минимизации процесса переобучения и максимизации вероятности правильного распознавания.

Литература

1. Капустий Б.Е. Оптимизация классификаторов в условиях малых выборок / Б.Е. Капустий, Б.П. Русын, В.А. Таянов // Автоматика и вычислительная техника. – 2006. – Вып. 5. – С. 25-32.
2. Капустий Б.Е. Математическая модель систем распознавания с малыми базами данных / Б.Е. Капустий, Б.П. Русын, В.А. Таянов // Проблемы управления и информатики. – 2007. – № 5. – С. 142-151.
3. [Електронний ресурс]. – Режим доступу : http:// www.ccas.ru/voron/teaching.html.
4. Ивахненко А.Г. Моделирование сложных систем по экспериментальным данным / А.Г. Ивахненко, Ю.П. Юрачковский. – М. : Радио и связь, 1987. – 120 с.
5. Feature Selection Using a Piecewise Linear Network / [Jiang Li, Michael T. Manry, Pramod L. Narasimha, and Changhua Yu] // IEEE Transactions on Neural Network. – September 2006. – Vol. 17, № 5. – P. 1101-1115.
6. Levner I. Automated Feature Extraction for Object Recognition / I. Levner // Proceedings of the Image and Vision Computing New Zealand Conference. – 2003. – P. 653-655.
7. Osborne M.R. A new approach to variable selection in least squares problems / M.R. Osborne, B. Presnell and B.A. Turlach // IMA Journal of Numerical Analysis. – 2000. – 20. – P. 389-404.
8. Mullin M. Complete cross-validation for nearest neighbor classifiers / M. Mullin, R Sukthankar // Proceedings of International Conference on Machine Learning. – 2000. – P. 639-646.
9. Капустій Б.О. Розподіл середньоквадратичних відстаней між об'єктами в просторі R^2 / Б.О. Капустій, Б.П. Русин, В.А. Таянов // Відбір і обробка інформації. – 2003. – Випуск 19(95). – С.110-114.
10. Капустій Б.О. Новый подход к определению вероятности правильного распознавания объектов множеств / Б.О. Капустій, Б.П. Русин, В.А. Таянов // УСиМ. – 2005. – № 2. – С. 8-13.

Б.О. Капустій, В.А. Таянов

Вплив розміру навчаючої вибірки на узагальнюючу властивість метричних алгоритмів класифікації

В роботі пропонується підхід, який забезпечує оцінку впливу зменшення розміру класів бази даних на рівень розпізнавання при застосуванні *k*NN метричних класифікаторів, а також дає можливість визначення на даній вибірці оптимального значення *k*. Проведене симулятивне моделювання результатів впливу зменшення навчаючої вибірки на результати розпізнавання. Отриманні результати можуть бути використані для подальшого формування навчаючої вибірки та її корекції.

B.O. Kapustiy, V.A. Tayanov

The Influence of the Training Set Size on the Generalized Ability of the Metrical Classifiers

In this paper the approach giving the estimate of the class size reduction influence on the recognition rate for the *k*NN classifiers has been proposed. The approach also gives the possibility to estimate the optimal *k* value of the nearest neighbours. The simulative modeling of the training set reduction influence on the recognition process results has been carried out. The obtained results can be used for the training set formation and its correction.

Статья поступила в редакцию 08.07.2009.